TÉCNICA

# EMPLEO DE **BIG DATA** EN LA **GESTIÓN** DE LAS **TELECOMUNICACIONES**

**Por:** Ing. Lieter Plasencia Moreno y Dra.C. Caridad Anías Calderón, ISPJAE.
*lieter.pm@gmail.com; cacha@tesla.cujae.edu.cu*

## RESUMEN

El alto desarrollo alcanzado en las Tecnologías de la Información a nivel global y el intenso uso por parte de los usuarios de las mismas han provocado el incremento de los volúmenes de datos que se transportan por las redes. La búsqueda de nuevos métodos para gestionar dichos datos conllevó al surgimiento del término Big Data, imponiéndose un nuevo paradigma en la gestión de los mismos.

En este artículo, se presenta cómo Big Data puede ser empleado en la optimización de la gestión de redes y servicios de las Telecomunicaciones, mediante el empleo de una arquitectura referencial de Big Data aplicable en este sector.

**Palabras clave:** Big Data, Gestión, Telecomunicaciones

## ABSTRACT

The high development reached in the Information Technologies worldwide and the intense deployment users make of these technologies, have caused the increase of data volumes that are transmitted through networks. The search of new methods to manage such data led to the appearance of the term Big Data, imposing a new paradigm in their management.

In this article, the author presents how Big Data can be used in the optimization of the management of Telecommunications networks and services, using a Big Data referential architecture applied in this sector.

**Key words:** Big Data, Management, Telecommunications

## Introducción

En los últimos años se ha apreciado una evolución acelerada de las Tecnologías de la Información y las Comunicaciones (TIC), destacándose el incremento de la interacción de los usuarios con las mismas, lo que ha provocado el tráfico de grandes cantidades de datos que no existían. Por otra parte, la digitalización de prácticamente todo en el nuevo mundo digital, ha dado lugar a nuevos tipos de datos en un amplio abanico de sectores, de los cuales muchos presentan una estructura que dificulta su procesamiento y almacenamiento.

De esta forma surge el término Big Data, que implica una nueva forma de gestionar el alto nivel de datos que existen y que se generan en la actualidad a nivel global, y aprovecharlos así en función de lograr las metas que se trazan las distintas empresas y organizaciones. Big Data no es más que la combinación de viejas y nuevas tecnologías que ayudan a obtener una mejor visión del procesamiento de la información [1]. También se puede conceptualizar como la capacidad de manejar un inmenso volumen de datos que se generan de forma caótica, los que a la velocidad y temporización correctas, permiten el análisis en tiempo real y la definición de las necesarias acciones asociadas.

Dado el nuevo panorama que presenta Big Data se han desarrollado nuevas herramientas y plataformas para el procesamiento de datos masivos que se generan en las redes desde diferentes fuentes y, de ellos, obtener información valiosa. La aplicación de Big Data en las telecomunicaciones contribuye al desarrollo de novedosos y potentes sistemas de gestión de redes y servicios. Algunas de las aplicaciones de Big Data en este sector son:

−Óptimo almacenamiento de datos masivos en la nube mediante almacenes virtualizados [2].

– Sistemas de monitoreo de redes capaces de procesar miles de datos en cuestiones de segundos.

– Técnicas de reubicación de datos en nodos de transmisión de gran velocidad y, con ello, la obtención de rutas más óptimas para el transporte de datos.

– Integración de los datos que circulan por las redes.

– Análisis de la información para la detección de fallos.

– Visualización de diferentes tipos de datos.

–Creación de *frameworks* para mejorar los servicios de comunicaciones móviles mediante la evaluación de la calidad de experiencia —*Quality of Experience* (QoE)— de los usuarios.

Este artículo abarca cómo Big Data puede ser aplicado en el sector de las telecomunicaciones. Para ello se emplea una arquitectura referencial de Big Data, la cual permite la obtención e integración de información de alto impacto en la gestión de las redes y los servicios de este sector en un caso de gestión de seguridad de una red.

## Propuesta de arquitectura

A partir de la investigación realizada, se elaboró la arquitectura referencial de Big Data para la gestión de las telecomunicaciones. Algunos de los principios que formaron la base de esta propuesta fueron:

- La necesidad de conocer qué datos son relevantes a los objetivos que se persiguen al emplear Big Data.

- La urgencia de requerir procesos de extracción, transformación y carga para garantizar la captura y almacenamiento de todo tipo de datos relevantes.

- La transformación de los datos que no presentan una estructura adecuada para su posterior análisis.

- El empleo de los sistemas aislados o de la nube para el almacenamiento de los datos, garantizándose que se almacenen todos aquellos que se capturen y se procesen.

- La determinación de las herramientas de análisis de datos a emplear según los objetivos que se persigan.

- El empleo de la virtualización, dadas las ventajas que proporciona.

- La necesidad de seguir un modelo de gestión de datos distribuido, puesto que las fuentes de las cuales estos se extraerán se encuentran geográficamente distribuidas.

- La garantía en todo momento de la seguridad de los datos, siendo este uno de los retos de la gestión de datos masivos.

En la figura 1 se muestra la arquitectura referencial de Big Data para la gestión de las telecomunicaciones propuesta. En el nivel más bajo de la arquitectura se encuentran las fuentes que generan grandes flujos de datos a diferentes velocidades y desde distintos puntos geográficos. En el segundo nivel aparecen los procesos de Extracción, Transformación y Carga —*Extraction, Transformation and Load* (ETL)— de los datos masivos.

El objetivo es extraer los datos de distintas fuentes y enviarlos a los repositorios donde se almacenan. Los procesos de transformación y carga pueden ocurrir de dos formas principales. En la primera, los datos son cargados inicialmente en las bases de datos que los almacenarán, dentro de estas, se hacen las transformaciones necesarias, lo que facilita que las herramientas de análisis de datos los procesen y entreguen la información de manera clara y comprendible. Mientras que en la segunda, los datos son transformados previamente al almacenamiento de los mismos.


**Figura 1.** Arquitectura referencial de Big Data para la gestión de las telecomunicaciones. Fuente: Elaboración propia.

En el tercer nivel de la arquitectura se considera el almacenamiento de datos masivos. Este nivel puede variar de una implementación a otra de la arquitectura, puesto que existen herramientas ETL que no solo transforman los datos sino que presentan espacios de almacenamiento para grandes volúmenes de información, no requiriéndose emplear bases de datos adicionales. Además, cada empresa u organización donde se aplique la arquitectura que se propone puede determinar, de acuerdo a los tipos de datos con los que va a trabajar, cómo almacenarlos.

En el cuarto nivel de la arquitectura se considera el análisis de datos, en el que se emplean herramientas que se encargan de obtener información de alto nivel de impacto, útil para la gestión de las redes y servicios de telecomunicaciones.





Ejemplos de estas herramientas son las de análisis predictivo de datos, algoritmos para establecer puntos de interrelación dentro de grandes volúmenes de datos y las de visualización que permitan representar información de interés sobre las redes y los servicios de las empresas de telecomunicaciones.

Finalmente, en el último nivel de la arquitectura referencial propuesta se encuentran las aplicaciones de gestión de las redes y servicios de telecomunicaciones, las cuales se ven optimizadas gracias al análisis de los datos masivos, como por ejemplo, para lograr la configuración eficiente de los dispositivos de interconexión de redes, la mejora en los servicios telefónicos y una mayor calidad de las ofertas a los clientes.

Se debe resaltar que el término de datos masivos se emplea en los procesos que ocurren en los niveles de Extracción, Transformación y Carga, Almacenamiento y Análisis de datos, pues solo después que los datos salen del nivel de Análisis, es que estos se consideran información relevante, es decir, información que puede ser aplicada en la gestión de redes y servicios de telecomunicaciones.

## Nivel de Extracción, Transformación y Carga de la arquitectura propuesta

El nivel de Extracción, Transformación y Carga de la arquitectura referencial de Big Data para la gestión de las telecomunicaciones se muestra en la figura 2. Los procesos realizados en dicho nivel son pilares al planificar y diseñar una infraestructura de manejo de datos que implique la integración de diferentes y variadas fuentes. Estos procesos son los responsables de recopilar la información de las fuentes de origen de datos adaptarla, filtrarla e integrarla en un repositorio digital, por ejemplo, una base de datos.

Los principales procesos a ejecutarse en el nivel de Extracción, Transformación y Carga se precisan en la figura 2 y sus características son:

– **Proceso de extracción:** se obtienen los datos de las fuentes de origen. Habitualmente, con el objetivo de evitar saturación en los servidores donde finalmente se almacenarán los datos, se suelen implementar repositorios intermedios, conocidos como bases de datos operacionales o almacenes de datos operacionales, que actúan de pasarelas entre las fuentes de datos y el sistema destino de la información.

– **Proceso de transformación:** cuando los datos proceden de distintas fuentes, lo común es que no coincidan en formato. Debido a esto, resulta imprescindible realizar tareas de transformación para, entre otros problemas, evitar duplicidades innecesarias de datos o que se establezcan grupos de datos que no presentan conexiones entre ellos. En este proceso se llevan los datos extraídos a una estructura lógica común necesaria para su procesamiento y análisis posterior.

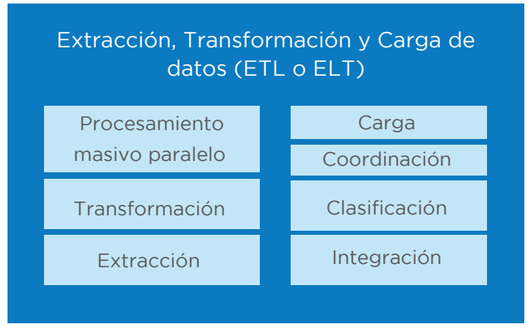


**Figura 2.** Procesos del nivel de Extracción, Transformación y Carga de la arquitectura propuesta. Fuente: Elaboración propia.

– **Proceso de clasificación:** permite la clasificación de los datos que se extraen en diferentes dimensiones para la simplificación de futuros procesamientos.

– **Proceso de integración:** armonización de datos de distintas fuentes y su integración en un grupo único antes de ser transformados y reducidos en un formato común.

– **Proceso de coordinación:** mantiene y controla a todos los demás procesos que se realizan en este nivel de la arquitectura.

– **Procesamiento Masivo Paralelo (MPP):** realiza la división de tareas para procesarlas al mismo tiempo y de forma aislada. Así, el sistema es más eficiente en el procesamiento de datos [3].

– **Proceso de carga:** se cargan los datos, ya estructurados en el formato deseado, en el sistema de almacenamiento destino donde posteriormente serán procesados y analizados [3].

## Nivel de almacenamiento de la arquitectura propuesta

El concepto de almacenes de datos se originó hace varias décadas. Inicialmente se concibió para que fuesen utilizados por usuarios que administraban sistemas operacionales que necesitaban almacenar información para apoyar la toma de decisiones. Con la llegada de Big Data el concepto de almacén de datos ha evolucionado; no obstante, los almacenes de datos tradicionales siguen siendo usados debido a que son eficientes en el análisis de datos operacionales antiguos.

Los almacenes de datos tradicionales soportan datos estructurados, están optimizados para propósitos específicos y generalmente son centralizados. Con la aparición de Big Data se ha pensado en almacenes de datos híbridos, en los que se encuentren tanto los datos estructurados como los no estructurados procesados por las herramientas ETL.

En la figura 3 se muestran las principales características que cualquier tecnología que se utilice para la implementación

del nivel de Almacenamiento de Datos de la arquitectura propuesta. Estas características son:

– **Replicación:** permite la redundancia de información en las bases de datos, con lo cual, si una base de datos deja de funcionar, la información se asegura pues se encuentra replicada en otras.

– **Balanceo de carga:** realiza la adecuada distribución de las bases de datos en múltiples servidores.

– **Escalabilidad horizontal:** permite que los datos se puedan almacenar en varios servidores. A mayor cantidad de información, más servidores se emplearán.

– **Sistemas distribuidos de ficheros:** opera con una red o clúster de servidores interconectados y configurados para trabajar con un sistema de ficheros lógico. El tamaño del sistema de ficheros puede variar, aumentar o disminuir, de acuerdo a las necesidades y sin afectar el rendimiento general del sistema.

– ***Sandboxing* o establecimiento de almacenes de datos temporales:** permite la creación de almacenes de datos temporales para la experimentación, el procesamiento y análisis de datos. Los datos que contienen son copiados desde la fuente donde se encuentran almacenados y libremente se puede escoger cómo se van a tratar los mismos y qué hacer con ellos, sin afectar los datos originales.

– **Filtros de datos:** permiten obtener datos específicos que se desean tratar o asegurar en el sistema de almacenamiento.

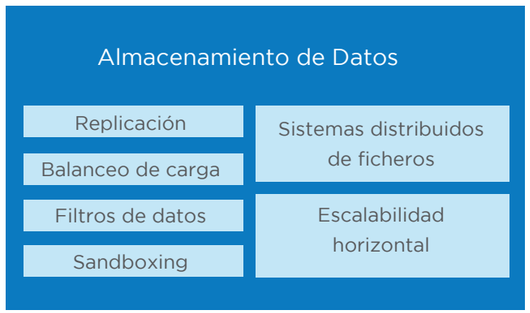


**Figura 3.** Características del nivel de Almacenamiento de datos de la arquitectura propuesta. Fuente: Elaboración propia.

Uno de los dilemas a los que se enfrentan muchas empresas es que no pueden costear la infraestructura física necesaria para almacenar grandes volúmenes de datos no estructurados. En la actualidad, muchos proveedores de almacenamiento de datos ofrecen soluciones para la nube que, como parte de su gama de productos, se comercializan entre los clientes como soluciones asequibles y accesibles. El almacenamiento en la nube permite que solo se necesite alquilar potentes servidores equipados con sofisticadas aplicaciones diseñadas especialmente para manejar grandes volúmenes de datos, a los que se puede acceder permanentemente. Existen varias ventajas del uso de la nube en entornos Big Data, entre ellas, la escalabilidad, la elasticidad, la utilización eficiente de recursos compartidos, la reducción de costos y la tolerancia a fallos.

## Nivel de Análisis de datos de la arquitectura propuesta

Teóricamente, una de las ventajas del empleo de Big Data es que mientras más datos se analicen, mayor será la amplitud de visiones que se puedan establecer en torno a los objetivos que persiga una organización o empresa. La primera pregunta que hay que formularse para seleccionar las herramientas de análisis de datos es: ¿qué o cuáles problemas se están tratando de resolver y en qué sector o área de la sociedad se encuentran enfrascados? También, en la selección de herramientas de análisis se debe tener en cuenta el nivel de complejidad del problema a resolver.

En la tabla 1 se recoge el empleo de algunas herramientas analíticas para varios tipos de análisis a realizar. En el Nivel de Análisis de datos de la arquitectura referencial para la gestión de redes y servicios de telecomunicaciones que se propone solo son de interés los dos primeros tipos de análisis.

Las analíticas básicas son utilizadas para explorar grandes volúmenes de datos. Permiten la división de estos en pequeños grupos que son más fáciles de analizar en tiempo real al posibilitar identificar en ellos anomalías e incidentes. En la gestión de las redes y servicios de telecomunicaciones, estas herramientas son de gran importancia para el monitoreo del desempeño de los dispositivos y los servicios de la red, la detección de anomalías y la visualización de las configuraciones.

Las analíticas avanzadas proveen algoritmos para análisis complejos de distintos tipos de datos, posibilitando su procesamiento y la obtención de patrones de los mismos para la predicción y prevención de eventos. Ejemplos del empleo de estas herramientas es la elaboración de modelos predictivos que faciliten a las empresas que brindan servicios de telecomunicaciones determinar comportamientos delictivos por parte de los usuarios, prevenir fallas y errores que puedan ocurrir, etc. También se pueden utilizar en el análisis de textos para extraer información valiosa o en el desarrollo de algoritmos que ayuden a la minería de datos. Es decir, se emplean en la obtención de información sobre los servicios o aplicaciones que se desean suministrar.





| TIPO DE ANÁLISIS | UTILIZACIÓN |
|---|---|
| Analíticas básicas | Selección y división de datos, reportes, visualizaciones simples y monitoreo básico |
| Analíticas avanzadas | Análisis complejos como modelos predictivos o técnicas de establecimiento de patrones |
| Analíticas operacionalizadas | Análisis de procesos de negocios |
| Analíticas monetizadas | Análisis monetarios |

**Tabla I.** Casos de uso de las herramientas analíticas. Fuente: [4].

En la figura 4, se muestran los principales procesos que deben realizarse en el Nivel de Análisis de datos de la arquitectura propuesta. Estos son:

–**Proceso de deducción de valor:** determina qué información es relevante a los objetivos que se persiguen.

–**Proceso de selección y división de datos:** se seleccionan y se dividen los altos volúmenes de datos en pequeños grupos para su análisis.

–**Proceso de determinación de patrones:** establece interrelaciones entre los grupos de datos obtenidos.

–**Proceso de visualización:** muestra en una interfaz gráfica el análisis que se realiza de los datos.

–**Proceso de diagnóstico y reporte:** permite obtener información útil resultante del análisis de los datos, aplicable a los objetivos que se persiguen.

–**Proceso de monitoreo:** permite conocer el desempeño y comportamiento del sistema.

En este punto se debe precisar que los procesos de los niveles de Extracción, Transformación y Carga, Almacenamiento de datos y Análisis de datos de la arquitectura propuesta son similares para diferentes aplicaciones de Big Data en la gestión de las telecomunicaciones. Sin embargo, los procesos del nivel de Gestión de redes y servicios de las telecomunicaciones dependerán de los objetivos específicos de gestión que se tengan.

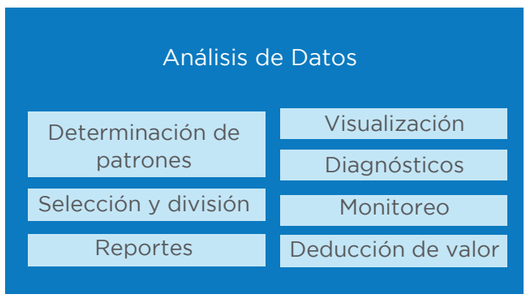


**Figura 4.** Procesos del nivel de Análisis de datos de la arquitectura propuesta. Fuente: Elaboración propia.

## Nivel de Gestión de redes y servicios de telecomunicaciones de la arquitectura propuesta

La gestión de las telecomunicaciones está compuesta por cinco áreas funcionales de la gestión: configuración, desempeño o prestaciones, fallos, seguridad y contabilidad. Algunos casos del empleo de Big Data en las áreas funcionales de la gestión de redes se explican en esta sección.

La obtención de información sobre las redes y sus servicios se ha visto beneficiada con el surgimiento de Big Data, que ha incorporado un amplio número de herramientas y oportunidades para el tratamiento de grandes cantidades de datos, estructurados y no estructurados. Las herramientas analíticas de datos masivos pueden ser empleadas en el análisis de transacciones financieras, en archivos de *logs* y en el tráfico de las redes, lo que permite identificar anomalías y actividades sospechosas y correlacionar coherentemente múltiples fuentes de datos.

Uno de los usos de Big Data es en la gestión de la seguridad de las redes. En un caso de estudio publicado, Zions Bancorporation anunció el empleo de *Hadoop* [5] (una de las principales herramientas de procesamiento de datos empleadas en entornos Big Data) y de otras herramientas de análisis inteligente que permiten trabajar con gran cantidad de datos en menor tiempo que con las herramientas de análisis de datos tradicionales según el estudio, se empleó de 20 minutos a 1 hora para analizar grandes volúmenes de datos utilizando las herramientas tradicionales, mientras que con *Hadoop* se realizó en 1 minuto aproximadamente. Además, se aumentó la seguridad de sus redes gracias al análisis efectivo de información proveniente de distintas fuentes como *firewalls*, dispositivos de redes, tráfico por la red, procesos de negocio y transacciones diarias.

Otro ejemplo de cómo Big Data puede ser empleado en la gestión de la seguridad de las redes es el trabajo realizado por HP Labs para identificar dispositivos infectados con *malware* (tipo de *software* malintencionado que daña los dispositivos) en las redes empresariales. Para ello se analizaron millones de datos de solicitudes del Protocolo de Transferencia de Hipertexto—*Hypertext Transfer Protocol* (HTTP)—, del Sistema de Nombres de Dominio —*Domain Name System* (DNS)— y de los sistemas de alerta de intrusos [6].

Gracias a la evolución de Big Data, se han logrado establecer mejores estrategias y métodos en la detección de Amenazas Persistentes Avanzadas —*Advanced Persistent Threat* (APT)— [7], que es uno de los problemas

más serios que enfrentan las empresas y organizaciones en cuanto a la seguridad de la información. Las APT, en contraste a otros tipos de *malware* como los troyanos y los gusanos, son agresores de las redes que trabajan en modo *low-and-slow*, es decir, *low* pues mantienen un perfil bajo en la red haciendo muy difícil su detección y *slow* porque están activos durante un largo periodo de tiempo.

Antes de Big Data, la detección de estas amenazas se basaba en la experiencia humana, lo que constituía una labor intensiva, difícil de generalizar y no escalable. Con el empleo de Big Data, el análisis de grandes grupos de datos ya no constituye un desafío, pues se ha logrado establecer un método que utiliza algoritmos de monitoreo para determinar prácticamente todos los posibles ataques en las redes.

Big Data puede traer grandes ventajas en el desempeño de las redes. Actualmente, se habla de redes más inteligentes, donde se logra la optimización de rutas para la entrega de miles de paquetes de información, así como la selección de los nodos o dispositivos de la red que poseen mayor cantidad de recursos disponibles para que por ellos circulen los mayores flujos de datos, aliviando así la carga de otros dispositivos y, de esta forma, eliminar o disminuir las colas de paquetes que puedan provocar pérdidas de datos y demoras en las redes.

Diversos estudios se refieren a la aplicación de Big Data en la mejora de la calidad de experiencia de los usuarios. Para ello, se utilizan técnicas de extracción y análisis de datos actualizados sobre las opiniones de millones de usuarios de diversos servicios, lo que permite que las empresas conozcan cómo sus servicios son aceptados y responder a las nuevas necesidades de los clientes. Otra forma de mejorar la QoE —*Quality of Experience*—, es optimizar la calidad de servicio —*Quality of Service* (QoS)— que brindan las redes y los servicios que se ofrecen, puesto que ambos conceptos se encuentran estrechamente relacionados.

Adicionalmente, se han trazado diversos acercamientos a la gestión de las redes y sus servicios en los nuevos entornos Big Data, entre ellos la gestión de la red basada en el valor —*Value-Based Network Management* (VBNM)—. La VBNM se sustenta en el análisis del comportamiento de los clientes y del consumo de los recursos de la red por parte de estos. También se basa en la extracción de información de los dispositivos la red, o sea, no solo considera los datos que circulan por ella, sino que tiene en cuenta la información brindada por los elementos de la misma, para lograr disminuir el consumo de recursos y el tiempo de retardo de la información, aumentar la eficiencia de los dispositivos, mejorar la configuración de estos, disminuir la congestión de la red y reubicar los recursos disponibles donde su utilización sea más productiva. [8]

IBM —*International Business Machines*—, una de las empresas más destacadas en tecnología y consultoría, opina que Big Data está hecho para la industria de las telecomunicaciones. Gracias al desarrollo de las redes y la proliferación de dispositivos inteligentes, los proveedores de servicios de telecomunicaciones tienen acceso a un gran cúmulo de información sobre los comportamientos y las preferencias de sus clientes. Actualmente, a nivel internacional, muchas empresas que brindan servicios de telecomunicaciones se encuentran enfrascadas en el desarrollo de alternativas para emplear Big Data en su gestión [9], una de las razones por la que se considera importante la investigación que en este artículo se presenta.

## Empleo de la arquitectura propuesta de Big Data en la gestión de seguridad de una red

La gestión de redes se basa en la planificación, instalación, supervisión y control de los elementos que forman una red para garantizar un nivel de servicio de acuerdo a un costo. Su objetivo es mejorar la disponibilidad, la relación calidad-costo y el rendimiento de las redes y servicios logrando una mayor productividad en la organización y un aumento de la satisfacción de los usuarios.

En esta sección se muestra cómo la arquitectura de Big Data propuesta puede ser empleada en un caso específico dentro de la gestión de redes. En dicho caso se integran para la gestión de la seguridad de la Red-Cujae y, en particular, en la detección de intrusiones, herramientas tradicionales empleadas en la seguridad de una red y herramientas empleadas en entornos Big Data.

Se escogió la Red-Cujae para la aplicación de la arquitectura referencial de Big Data que se propone, teniendo en cuenta sus dimensiones, características técnicas, volumen de servicios y usuarios. En esta red es necesario, como en muchas otras, una mejor integración de los datos para optimizar los procesos de gestión y sus servicios.

No obstante, para lograr en la Red-Cujae una adecuada implementación de un sistema de gestión de datos masivos como Big Data, sería necesario realizar modificaciones y actualizaciones, tanto de hardware como de software, teniendo en cuenta que esta tecnología demanda muchos recursos. En particular, se requiere una gran capacidad de procesamiento y almacenamiento en los servidores, dados los volúmenes de datos que se necesitan tratar, los cuales están relacionados con la forma en que los usuarios perciben los servicios, el desempeño de la red y de los dispositivos de interconexión y la detección de fallas y amenazas.

Los principales puntos de interés para aplicar la arquitectura propuesta a la gestión de la seguridad de la red son:





– Definir las principales fuentes de dónde serán extraídos los datos.

– Determinar las herramientas necesarias para la extracción, transformación y carga de los datos, desde las fuentes que los generan hasta los sistemas de almacenamiento.

– Definir un sistema de detección de intrusiones basado en herramientas de gestión de seguridad de la red y entornos de gestión de datos masivos para la correcta detección de anomalías en la red.

– Establecer un sistema de almacenamiento para (archivar) los datos capturados.

– Definir los procesos de análisis y las herramientas necesarias para ejecutarlos. Esto permite mejorar la gestión de la seguridad de la red.

Es importante destacar que el último nivel de la arquitectura debe considerarse dentro del objetivo específico en el que se desea emplear la arquitectura propuesta. En este caso será en la gestión de seguridad de la Red-Cujae y específicamente en la detección de intrusiones. El esquema del caso que se explica se muestra en la figura 5.

Para el establecimiento del sistema de detección de intrusiones en la Red-Cujae se empleará *Snort* [10] y *Hadoop*. *Snort* es un analizador de paquetes y detector de intrusos —*Intrusion Detection System* (IDS)— que ofrece capacidades de almacenamiento tanto en archivos de texto como en bases de datos *open source*. Implementa un motor de detección de ataques y monitoreo de puertos que registra, alerta y responde ante las anomalías previamente definidas. Posibilita, entre otras funciones, la observación del funcionamiento y el tráfico de la red en tiempo real.

Por su parte, *Hadoop* es una herramienta de código abierto con un alto desempeño en el procesamiento de datos masivos, la cual fue seleccionada para la extracción, transformación y carga de los datos que captura *Snort*. Cuenta con distintos componentes que se encargan de las funciones ETL de datos no estructurados, que en muchas herramientas no existen, permitiendo la extracción masiva de datos en cuestiones de segundos.

La mayoría de los sistemas de detección de intrusos identifican rápidamente ataques a partir de una serie de reglas. Los paquetes entrantes son analizados y comparados con las reglas definidas y si no cumplen con las reglas, entonces acciones especificadas se realizarán. Es obvio que a mayor cantidad de reglas definidas, mayor número de amenazas se podrán identificar. La mayor desventaja de los sistemas de detección de intrusos es que no son capaces de identificar ataques desconocidos, es decir, eventos que no se encuentran definidos en sus reglas.

En el caso analizado, la principal fuente de la cual se extraerán los datos hacia *Hadoop* será de *Snort*. Dicha herramienta presenta varios modos de ejecución. Uno de estos modos es el *Packet Logger*, en el cual *Snort* analiza el tráfico de la red, captura los paquetes de interés y los almacena temporalmente en una unidad de almacenamiento. Aquí se guardarán en el Sistema Distribuido de Archivos de *Hadoop*, —*Hadoop Distributed File System* (HDFS)—.

Los principales datos de interés para la aplicación que se analiza serán los paquetes que circulan desde o hacia los distintos nodos que se encuentran distribuidos en la red. Estos datos son generados por los dispositivos de la red, los usuarios internos y externos, las aplicaciones, etc.

Existen ataques que se caracterizan por el envío de un gran número de paquetes hacia un dispositivo como son los ataques del Protocolo de Mensajes de Control de Internet —*Internet Control Message Protocol* (ICMP)—, los *pings* de la muerte, los ataques *smurf*, los ataques del Protocolo de Datagramas de Usuario —*User Datagram Protocol* (UDP)—, entre otros. En la detección de este tipo de amenazas se centra principalmente este caso de uso.

*Hadoop* es capaz, en la medida que va extrayendo los datos de *Snort*, de identificar y clasificar desde y hacia donde están dirigidos los paquetes y determinar la cantidad que

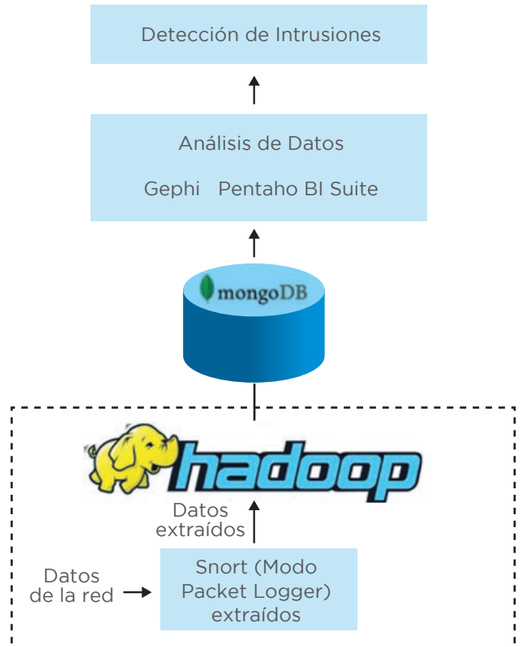


**Figura 5.** Esquema del sistema Big Data para la de detección de intrusiones. Fuente: Elaboración propia.

estos son, empleando para ello las banderas y los campos del protocolo IP —*Internet Protocol*— de los paquetes como la dirección fuente y destino y el número de puerto. El procesamiento de los datos dentro de *Hadoop* se muestra en la figura 6.

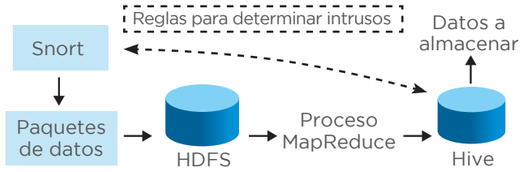


**Figura 6.** Procesamiento de los datos en Hadoop. Fuente: Elaboración propia.

Además, *Hadoop* realiza procesos de mapeo y reducción para eliminar los paquetes que presentan la misma información, así como la entrega la información en un formato adecuado para su posterior análisis. Los datos son enviados a Hive [5], componente de *Hadoop*, el cual a partir de las reglas implementadas en *Snort*, es capaz de identificar cuando existe un ataque hacia un nodo o dispositivo, marcando los paquetes identificados para reconocer que son un ataque. Hive tendrá las reglas definidas en *Snort* implementadas en su lenguaje de consulta HiveQL. Dada la facilidad del lenguaje de programación de Hive es posible elaborar nuevas reglas que pueden ser añadidas a *Snort* para la detección de anomalías.

La principal ventaja que presenta la integración de *Hadoop* y *Snort* es que las reglas se pueden establecer en *Snort* previo y posterior al análisis de los datos. Generalmente, *Snort* analiza los paquetes para determinar posibles amenazas, guiándose por las reglas que previamente fueron establecidas. Esto puede generar un alto número de falsas alarmas ante la llegada de paquetes desconocidos. Con el empleo de *Hadoop* esto puede mejorarse, ya que como antes se dijo este es capaz de analizar los paquetes y elaborar nuevas reglas a partir de las amenazas que se encuentran en los datos analizados. Estas nuevas reglas son incorporadas en Hive incrementándose la capacidad de detección de amenazas.

Para el almacenamiento de todos los datos de la red, cumpliendo con la implementación de la arquitectura propuesta en la Red-Cujae, se recomienda utilizar como base de datos principal a MongoDB, la cual es no relacional, orientada a documentos y ampliamente utilizada a nivel internacional. El objetivo es que después de analizados los datos en *Hadoop*, estos sean almacenados en un sistema que sea escalable y altamente disponible para su uso posterior. Es importante precisar que en MongoDB se almacenarán todos los datos que sean necesarios para la gestión de la Red-Cujae, no solo los requeridos para la gestión de seguridad.

Las principales razones que se tuvieron en cuenta para escoger a MongoDB como base de datos son:

– Alto nivel de integración con *Hadoop* en escenarios Big Data, permitiendo el almacenamiento de gran variedad de datos.

– Simple instalación

– Alto nivel de escalabilidad

– Bajos costos de implementación

– Alta adopción a nivel internacional

Las dos herramientas seleccionadas para el análisis de los datos almacenados en MongoDB fueron: Pentaho BI Suite [11] y Gephi [12]. Pentaho constituye un conjunto de programas *open source* que incluyen los componentes principales requeridos para implementar soluciones basadas en procesos. Posee una *Web* organizada en productos o componentes de reporte, análisis, minería de datos y *dashboards*, y es altamente utilizado para el acceso, integración, visualización y exploración de todo tipo de datos que puedan impactar en los negocios. Pentaho fue seleccionada para la aplicación de la arquitectura propuesta ya que soporta los principales procesos de análisis de datos que se desean implementar: minería de datos y análisis predictivo de fallas y amenazas.

Gephi se seleccionó principalmente por la alta capacidad de visualización de redes que provee. Es una herramienta *open source*, creada para facilitar que el usuario explore la red, la visualice y realice análisis en tiempo real. Además, por sus características es altamente aplicable a la gestión de los servicios de una red.

El empleo de estas herramientas de análisis representa grandes ventajas para la gestión de la seguridad de una red y, en particular, para la detección de intrusiones. Primeramente, mediante Gephi se puede realizar un esquema que visualice la red y sus elementos, lo que ayuda a determinar las principales zonas de riesgo, los elementos de la red más vulnerables y dónde han ocurrido mayor cantidad de amenazas. Esto facilita que se puedan llevar a cabo las acciones necesarias para optimizar la detección de intrusiones.

Con el empleo de Pentaho se pueden desarrollar diagramas donde se muestren los resultados del análisis del tráfico de la red, determinándose aquellos parámetros que más influyan en la optimización de la seguridad de la red y la gestión de la misma. También permite realizar el análisis predictivo de amenazas a partir de los datos que se encuentran almacenados en MongoDB y de los obtenidos en tiempo real.





## Conclusiones

La tecnología Big Data tiene un alto nivel de aplicación en el sector de las telecomunicaciones. Su arquitectura referencial para la gestión de las telecomunicaciones que se propone es de interés para la gestión de redes y servicios de distintos escenarios.

Mediante este sistema de detección de intrusiones se pudo aplicar la arquitectura referencial de Big Data para la gestión de las Telecomunicaciones en un caso real que puede demostrar que la arquitectura propuesta de Big Data es aplicable a la gestión de redes y que puede ser implementada con las tecnologías existentes de software libre y código abierto.

## Referencias bibliográficas

[1] Hurtwitz, J.: Big Data for Dummies, John Wiley & Sons. New Jersey, 2013.

[2] Liu, Z.: "A Domain Scientific Data Cloud Based on Virtual Dataspaces", Parallel and Distributed Processing Symposium Workshops PhD & Forum (IPDPSW) IEEE 26th International, pp. 2176-2182, 2012.

[3] Kasibhotla, D.: Introduction to Massively Parallel Processing (MPP) Database, URL: https://dwarehouse.wordpress.com/2012/12/28/ introduction -to-massively-parallel-processing-mpp-database. Fecha de consulta: 4 de febrero de 2015.

[4] Schoenborn, B.: Big Data Analytics Infrastructure for Dummies, John Wiley & Sons, New Jersey, 2014.

[5] White, T.: Hadoop: The Definitive Guide, O´Reilly Media, Sebastopol, 2011.

[6] Big Data Analytics for Security Intelligence, Cloud Security Alliance White Paper, 2013.

[7] Giura, P. y W. Wang: Using Large Scale Distributed Computing to Unveil Advanced Persistent Threats, ASE, 2012.

[8] Arias, J.: Value-Based Network Management for Telecoms, 2015.

[9] Fox, B.; R. Dam y R. Shockley: "Analytics: el uso de Big Data en el mundo real aplicado a las Telecomunicaciones", IBM Global Business Services, Business Analytics and Optimization, 2013.

[10] Prathibha, P.G. y E.D. Dileesh: "Design of a Hybrid Intrusion Detection System using Snort and Hadoop", International Journal of Computer Applications, Vol.73, No.10, 2013.

[11] Goodman, N.: Pentaho Data Integration, Bayon Technologies White Paper, 2009.

[12] Amat, C.B.: "Análisis de redes y visualización con Gephi", REDES- Revista Hispana para el análisis de redes sociales, 2014.

(Artículo recibido en noviembre de 2015 y aprobado en febrero de 2016)